基因預測軟體在水稻白葉枯病抗性基因的應用性評估及分析平台建立

呂椿棠、呂秀英*、陳政道
行政院農委會農業試驗所作物組

摘要

利用基因預測軟體在含有水稻白葉枯病抗性基因(Xa)結構特徵的大片段序列集合中，找出最有可能的 Xa 候選基因，再以基因定位試驗進行鑑別，可加速水稻抗病基因之發現。為找出適用於目前已知序列之 Xa 基因結構型態的基因預測軟體，以利未來之 Xa 基因預測工作，本研究蒐集以馬可夫模式為基礎所開發的 5 種基因預測軟體：GlimmerM、GlimmerR、RiceHMM、GeneMark.hmm 及 GENSCAN，以 Xa1、Xa21、Xa26 及 Xa27 四個基因已知的 18 條 DNA 序列為測試資料，分別在核苷酸與外顯子之層次上評估其預測的靈敏性和專一性。分析結果顯示在核苷酸與外顯子兩個層次上，皆以 GeneMark.hmm 軟體之準確度最高，其次為 GENSCAN 軟體。另外，為使基因預測結果之展現及比較能更一目了然，亦以 Bioperl 模組程式設計出一套基因模式化視覺平台，提供將所預測之序列結構以不同顏色圖形瀏覽。據此建立了適用於水稻白葉枯病抗性基因預測的簡易分析程序與可應用分析平台。

關鍵詞： 水稻白葉枯病抗性基因、馬可夫模式、基因預測、基因模式化視覺平台。

* 通信作者, iying@wufeng.tari.gov.tw
投稿日期：2006 年 5 月 27 日
接受日期：2006 年 7 月 24 日
作物、環境與生物資訊 4:246-258 (2007)
Crop, Environment & Bioinformatics 4:246-258 (2007)
189 Chung-Cheng Rd., Wufeng, Taichung Hsien 41301, Taiwan ROC

Evaluation of Gene Prediction Programs in Finding Rice Bacterial Blight Resistance Genes and Establishment of the Analytical Platform

Chun-Tang Lu, Hsiu-Ying Lu* and Cheng-Tao Chen

Crop Science Division, Agricultural Research Institute, Council of Agriculture, Wufeng, Taichung Hsien 41301, Taiwan ROC

ABSTRACT

To accelerate the gene finding of rice resistant to disease, it is essential to use gene prediction programs to find the most probable candidate, rice bacterial blight resistance gene (Xa), encoded in large rice sequence sets with Xa gene structure, and then perform gene mapping to experimentally verify the prediction result. We believe a more realistic evaluation of various gene prediction tools on currently data sets with Xa gene structure would help our future work in Xa gene prediction. Thus, we collected five gene prediction programs based on Markov models, i.e., GlimmerM, GlimmerR, RiceHMM, GeneMark.hmm and GENSCAN, and used a total of 18 complete rice DNA sequences of bacterial blight resistance genes, i.e., Xa1, Xa21, Xa26 and Xa27, as test data. Sensitivity and specificity measures were employed to evaluate their gene prediction accuracy at both nucleotide and exon levels. Results showed that GeneMark.hmm was the most accurate in both nucleotide and exon prediction, and followed by GlimmerM. We also used Bioperl modules to design a gene modeling visualization platform that enabled a clear comparative display of the sequence structure

predicted by the five programs. The graphic output provided a number of pre-defined color schemes that assisted in visual identification of the gene prediction differences among the five prediction programs. A simple analytical procedure and an analytical platform were thus established for gene prediction in rice bacterial blight resistance genes.

Key words: Rice bacterial blight resistance genes, Markov models, Gene prediction, Gene modeling visualization platform.

前言

人類基因體計畫(Human Genome Project, HGP）與國際水稻基因體定序計畫(International Rice Genome Sequencing Project, IRGSP)分別於2000及2002年相繼宣佈完成定序工作，正式揭示後基因體時代的來臨(IHGSC 2001, Goff *et al.* 2002)。在此之前的研究，通常是先確認基因在染色體上的位置後再定序出該基因的序列，其所需時間與費用較高；而今基因體序列已被解讀且公開查詢，該如何利用生物資訊學(bioinformatics)方法，快速且有效地對此基因體序列資料庫進行未知基因序列的預測工作，成為當前的挑戰。基因是生物體 DNA 序列上的一段功能區，真核生物(eukaryotic)的基因通常可簡單分為啟動子(promoter)、基因上游區域非編碼區(5'untranslated region, 5'UTR)、外顯子(exon)、內含子(intron)、基因下游區域非編碼區(3'UTR)等部分。此外，基因與基因間則存有基因間序列(intergenic sequence)，其中只有外顯子才攜帶有產生蛋白質的訊息，此種結構遠較原核生物(prokayotic)來得複雜，且預測的複雜度也較高。而利用基因預測軟體進行基因辨識，最主要的任務就是從 DNA 序列，找出一個基因所包含之外顯子的起始碼(start codon)、終止碼(stop codon)及剪接點(splicing site)的位置，進而將同一基因所包含的外顯子拼湊出來，組成一個完整的基因構造，且正確轉譯為胺基酸序列，最後希望能以此為基礎進行蛋白質二級與三級結構之模擬等研究(Burge and Karlin 1997, Mathé *et al.* 2002)。

第一個被發展出來的基因預測軟體之計算基礎，是根據編碼區(codon)和非編碼區(noncodon)之間的組成差異，來比較不同讀取框架(frame shift)內的密碼子偏好性(codon usage)(Staden and McLachlan 1982)。近年來由於分子序列數量的急速成長，利用基因預測軟體進行新基因預測，已成為實驗室外快速發現新基因的一種手段，因而新的基因預測軟體不斷被推陳出新。現今已開發的基因預測軟體相當多，大致上可將它們區分為兩類：一為從頭預測(*ab initio*)方法，乃根據統計分析原理，利用已知的基因結構組成訓練資料集(training data)，針對編碼區與非編碼區的生物訊號特性對預測模式進行訓練，並根據預測結果調校參數，以提升預測之準確度，此類軟體如 GENSCAN (Burge and Karlin 1997)、HMMGene (Krogh 1997）、Glimmer (Salzberg *et al.* 1998, Delcher *et al.* 1999)、RiceHMM (Sakata *et al.* 1999）、GeneMark (Borodovsky and McIninch 1993, Lomsadze *et al.* 2005)等。這些大都以馬可夫鏈(Markov chain)及其變種模式為基礎，諸如內插式馬可夫模式(Interpolated Markov models, IMM)、隱匿馬可夫模式(Hidden Markov model, HMM）、隱匿半馬可夫模式(Hidden semi-Markov model, HSMM)等。另一類為相似性(similarity-based)方法，乃透過分析未知序列與資料庫中已知的基因序列間之相似性來推斷其基因的存在，此類軟體如 GENEWISE (Birney and Durbin 1997)、GENOMESCAN (Yeh *et al.* 2001)等。

基因預測方法之開發研究上，目前仍以從頭預測方法應用最多，其中 GeneMark 和 Glimmer 兩種預測軟體，最初是以預測原核生物的基因為目的而開發，隨後為因應真核生物基因預測之需要，再根據真核生物的基

因結構特性加以修改程式，讓預測軟體能適用更廣泛的生物體 (Borodovsky and McIninch 1993, Delcher *et al.* 1999, Lomsadze *et al.* 2005)。RiceHMM 是由 RGP (Rice Genome Research Program, http://rgp.dna.affrc.go.jp/E/rgp/rgpintro.html) 機構針對水稻表現序列標幟 (expressed sequence tags, EST)所發展的預測軟體(Sakata *et al.* 1999)。至於 GENSCAN 預測軟體在發展之初，便以人類基因體序列之基因預測為目的，其基因預測準確度相當高，故常被拿來做為基因預測軟體比較的基準 (Burest and Guigó 1996, Burge and Karlin 1997, Pedersen and Hein 2003)。然而，不同軟體預測基因結構的效率及準確度 (accuracy) 各有不同，所得分析結果有別 (Mathé *et al.* 2002, Wang *et al.* 2004)。一個高準確度的基因預測軟體，必須先經過一批精心挑選的訓練資料集建立基因結構特徵的統計學模型，然後利用這些模型對未知 DNA 序列的基因結構進行預測。回顧近年之文獻，用來評估真核生物之基因預測軟體準確度的測試材料，大都包括人類(human)、大鼠(rat)、小鼠(mouse)、果蠅(*Drosophila*)與阿拉伯芥(*Arabidopsis*)等之基因體序列(Burge and Karlin 1998, Pavy *et al.* 1999, Rogic *et al.* 2001, Rogic *et al.* 2002, Pedersen and Hein 2003, Stanke and Waack 2003, Brejová *et al.* 2005)，且基因預測準確度所使用的評估指標亦都包括 Burest and Guigó (1996)的靈敏性 (sensitivity, Sn)與專一性(specificity, Sp)之雙重指標方法。

白葉枯病是水稻最嚴重的細菌性病害，由黃單孢桿菌 (*Xanthomonas oryzae* pv. *oryzae*, 簡稱 Xoo)所引起，目前栽培管理上並無有效之防治方法，嚴重時可造成水稻減產約 20-60%，因此育成抗白葉枯病品種視為唯一重要防治方法(Adhikari 1999)。水稻基因體定序雖已於 2002 年完成初稿，但目前在白葉枯病抗性基因(Xa)序列定位之研究成果上，仍僅限 Xa21 (Song *et al.* 1995)、Xa1 (Yoshimura *et al.* 1998, Yu *et al.* 2003)、Xa26 (Sun *et al.* 2004, Sun *et al.* 2006)與 Xa27 (Gu *et al.* 2005)等 4 個基因已有完整的 DNA 序列，其他 Xa 基因的序列則尚不可知。自國際水稻基因體定序計畫完成後，序列數量的成長相當快速，部分基因預測軟體也因應此需求而加入水稻的基因序列為訓練資料，以進行模式修正及改版。然而在這些使用介面已提供水稻基因序列作為預測序列選項的軟體，都是廣泛針對所有水稻基因序列進行測試，其預測效果是否仍適用於特定基因則尚不可知。因此本研究擬藉由生物資訊學方法，評估以馬可夫模式所開發的 5 個基因預測軟體，利用已定序之水稻白葉枯病抗性基因序列為測試數據，藉由靈敏性與專一性評估指標之估算，並建立其分析平台。未來將以預測準確度高的基因預測軟體在含有水稻 Xa 結構特徵的大片段序列集合中，找出最有可能的 Xa 候選基因，再以基因定位試驗進行鑑別，可加速水稻抗病基因之發現。

材料與方法

一、相關基因序列的取得

本研究之水稻白葉枯病抗性基因序列的取得，首先於 Oryzabase 網站 (http://www.shigen.nig.ac.jp/rice/oryzabase/top/top.jsp)以其提供的性狀基因(trait genes)之類別或符號功能，配合水稻白葉枯病學名"*Xanthomonas campestris* pv. *Oryzae*" 與罹病名稱"bacterial blight "進行交叉比對，確認水稻白葉枯病抗性基因之基因符號 (gene symbol)，分別以 Xa#表示顯性基因，xa#為隱性基因，目前已命名 32 個基因。經以此基因符號查詢已登錄的 DNA 序列資料庫，並逐一檢視查詢所得註解內容，篩選出含有 Xa 基因編號註解的序列，再連結至 NCBI 網站(http://www.ncbi.nlm.nih.gov/) 檢視是否為已被完全定序的基因。結果共蒐

集到 Xa1、Xa21、Xa26 與 Xa27 等 4 個顯性基因，分別分布於 4、6、10、11 號染色體上，其中 Xa21 基因家族包括 7 個基因序列、Xa26 基因家族包括 4 個基因序列，如 Table 1 所示。Xa21 基因家族除 Xa21B 外皆來自非洲的長花藥野生稻，其他基因大都來自印度型栽培稻。進行基因預測軟體之評估，若能蒐集較多已知結構的受測序列，則對於基因的實際與預測位置之評估，所獲得結果始能客觀。Xa1 基因有來自不同發表時間的兩條序列(各以 Xa1-1999、Xa1-2004 表示)，Xa21C 基因序列其實內含 Xa21B 與 Xa21C 兩段 DNA 序列，而可將其分割成兩個各具完整的基因序列(此處各以 Xa21B*和 Xa21C*標示，已與原有的 Xa21B 和 Xa21C 基因序列有所區別)，而 Xa26 基因家族另有 2004 年發表之 Xa26 DNA 序列(以 Xa26b*表示)，Xa27 基因有兩個不同水稻品種來源的 DNA 序列(各以 Xa27-IR24 和 Xa27-IRBB27 標示)，這些加上原有基因序列，共 18 個 DNA 序列，其 DNA 長度、CDS(codon sequences)範圍與長度、外顯子數目等資料整理如 Table 1。各基因 CDS 的長度自 342 bp 到 5409 bp 不等，外顯子個數自 1 個到 3 個不等，於是本研究利用這些已知 Xa 基因序列，供作基因預測軟體評估之測試數據。

二、相關分析軟體蒐集

馬可夫鏈是一種隨機過程。所謂第一階馬可夫鏈(1st –order Markov chain)，是指隨機變數在某一階段的機率分布只與該變數在前階段所處的狀態有關，而與之前歷史狀態無關。然而，DNA 序列有其結構上的特性，每 3 個核苷酸組成一個胺基酸的密碼子，由於每種胺基酸在特定基因上具有密碼子偏好性(codon usage)，因此馬可夫鏈可因應不同狀態而產生不同階層的馬可夫模式(kst –order Markov models, kst MM)(Burge and Karlin 1998, Salzberg *et al.* 1998)。隱匿馬可夫模式是近十幾年來被廣泛發展的時間序列模型，最早在語音識別(speech recognition)(Rabiner 1989)方面運用，後來與機器學習(machine learning)技術結合，廣泛應用於生物資訊學領域，如 DNA 編碼區之辨識、蛋白質超級家族(protein super family)模型之建構與分析等(Krogh *et al.* 1994, Henderson *et al.*1997)。本研究藉由文獻收集及網路搜尋，共篩選出 GlimmerM、GlimmerR、RiceHMM、GeneMark.hmm 與 GENSCAN 等 5 種以馬可夫模式為基礎的基因預測軟體，除 GlimmerM 與 GENSCAN 外其餘 3 個皆已經過水稻序列訓練且為免費軟體，這些軟體的基礎方法、分析平台、版本與來源，示如 Table 2。其中，除了 GlimmerM 必須下載至 Linux 作業環境的個人電腦上執行外，其餘 4 個軟體都能直接透過 web 進行線上分析。

Glimmer 為 Salzberg *et al.* (1998)針對細菌等微生物所發展的基因預測軟體，針對原核生物中只有一個長的開放讀取框架(open reading frame, ORF)結構下，結合 0 至 8 階 MM 模式，並針對各模式之預測冪(predictive power)與以加權的 IMM 方法，找出正確的編碼區。本研究依 TIGR 網站(http://www.tigr.org/software/genefinding.shtml)上所提供之服務內容，及其所適用的電腦分析平台區分為 GlimmerM (linux 作業平台)與 GlimmerR (web 線上操作)(Delcher *et al.* 1999)。GeneMark.hmm 軟體則由 Borodovsky and McIninch (1993)針對真核生物而開發，根據 DNA 序列中的編碼區與非編碼區結構特性，對編碼區採用 5 階 MM 而非編碼區則採用同質 MM (homogeneous MM)模式來建構蛋白質編碼區；該軟體後來並導入 HMM 架構於預測模式中(Lukashin and Borodovsky 1998)，近期更導入自動化的訓練與基因預測演算法，增加對於真核生物的基因預測，並依物種類別區分為 GeneMark、GeneMark.hmm、GeneMarkS 等名稱，本研究採用針對真核生

Table 1. Information summary of known sequences in rice Xa genes from Oryzabase and NCBI databases. (update time: March 30, 2007)

Label [y]	Gene symbol	Chr. no.	Submitted year	Organism	Accession no. in NCBI	DNA length(bp)	CDS range	CDS length(bp)	Exon no.
Xa1-1999	Xa1	4	1999	*Oryza sativa* (indica cultivar-group)	AB002266	5910	113..5521	5409	1
Xa1-2004	Xa1	10	2004	*Oryza sativa* (japonica cultivar-group)	NM_194685	4206	1..4206	4206	1
Xa21-A1	Xa21-A1	11	1997	*Oryza longistaminata*	U72725	8416	4771..7384,7676..8052	2991	2
Xa21-A2	Xa21-A2	11	1998	*Oryza longistaminata*	U72727	5940	2151..4764,4872..5248	2991	2
Xa21-B	Xa21-B	11	1995	*Oryza sativa* (indica cultivar-group)	U37133	3921	1..2677,3521..3921	3078	2
Xa21-B*	Xa21-B	11	1998	*Oryza longistaminata*	U72723	9940	5213..7889,8732..9132	3078	2
Xa21-C	Xa21-C	11	1998	*Oryza longistaminata*	U72723	19639	15118..17720,17827..18201	2978	2
Xa21-C*	Xa21-C	11	1998	*Oryza longistaminata*	U72723	9699	5178..7780,7887..8261	2978	2
Xa21-D	Xa21-D	11	1997	*Oryza longistaminata*	U72726	13341	2367..4205	1839	1
Xa21-E	Xa21-E	11	1997	*Oryza sativa* (indica cultivar-group)	U72724	9424	2819..5260	2442	1
Xa21-F	Xa21-F	11	1998	*Oryza longistaminata*	U72728	7204	1683..4352,5147..5547	3071	2
Xa26a	Xa26MRKa	11	2006	*Oryza sativa* (indica cultivar)	DQ355952	3401	1..2911,3016..3401	3297	2
Xa26b*	Xa26	11	2004	*Oryza sativa* (indica cultivar)	AY364476	6000	1542..4476,4582..4958	3312	2
Xa26b	Xa26MRKb	11	2006	*Oryza sativa* (indica cultivar)	DQ355952	3631	49..2983,3089..3631	3312	2
Xa26c	Xa26MRKc	11	2006	*Oryza sativa* (indica cultivar)	DQ355952	3456	1..2926,3041..3456	3342	2
Xa26d	Xa26MRKd	11	2006	*Oryza sativa* (indica cultivar)	DQ355952	11396	1..1403,7285..8819, 11008..11396	3327	3
Xa27-IR24	Xa27	6	2005	*Oryza sativa* (indica cultivar) IR24	AY986491	2393	1590..1931	342	1
Xa27-IRBB27	Xa27	6	2005	*Oryza sativa* (indica cultivar) IRBB27	AY986492	2361	1558..1899	342	1

[y] Xa1-1999 and Xa1-2004 were submitted in 1999 and 2004, respectively; Xa21-B* and Xa21-C* were partitioned from Xa21-C; Xa26b* was submitted in 2004; Xa27-IR24 and Xa27-IRBB27 were sequenced from IR24 and IRBB27 rice varieties, respectively.

Table 2. Information summary of gene prediction programs.

Program	Probabilistic method	Analytical platform	Trained on rice data	Source	Author	Current version
GlimmerM	Interpolated Markov model	Linux-PC	no	http://www.tigr.org/software/glimmerm/	Delcher *et al.* (1999)	3.0
GlimmerR	Interpolated Markov model	Web	yes	http://www.tigr.org/software/glimmerm/	Delcher *et al.* (1999)	1.0
RiceHMM	Hidden Markov model	Web	yes	http://rgp.dna.affrc.go.jp/RiceHMM/	Sakata *et al.* (1999)	2.0
GeneMark.hmm	Hidden semi-Markov model	Web	yes	http://opal.biology.gatech.edu/GeneMark/	Lomsadze *et al.* (2005)	3.3
GENSCAN	Fifth-order Markov model	Web	no	http://genome.dkfz-heidelberg.de/cgi-bin/GENSCAN/genscan.cgi	Burge and Karlin (1997)	1.0

物的 GeneMark.hmm 作為分析軟體(Lomsadze *et al.* 2005)。RiceHMM 是由 RGP 機構所開發，採用 HMM 的機率模式，為針對水稻 EST 序列所發展的預測軟體(Sakata *et al.* 1999)。GENSCAN 軟體由 Burge and Karlin (1997)所開發，發展之初便以人類基因體序列為材料，使用 3 週期 5 階 MM (three-periodic 5th-order MM)以預測 DNA 編碼區。

三、序列資料分析與軟體效益評估

各類型分析軟體之效益評估，除了要求操作介面必須簡便外，分析結果的準確度當屬重要因素，因此本研究採用 Burset and Guigó (1996)針對基因預測所定義之靈敏性與專一性兩種指標，來評估各種基因預測軟體對水稻白葉枯病抗性基因之預測準確度。其評估方法區分核苷酸與外顯子兩種不同層次，說明如下：

1. 核苷酸序列層次

針對編碼區與非編碼區之預測，考慮四種可能預測結果(Fig. 1a)：TP (true positives)為正確預測編碼區的核苷酸數目、TN (true negatives)為正確預測非編碼區的核苷酸數目、FN (false negatives)為錯誤預測非編碼區的核苷酸數目、FP (false positives)為錯誤預測編碼區的核苷酸數目。針對每個核苷酸為單位進行計算，以靈敏性表示實際編碼區內之編碼核苷酸被預測到的比率，而專一性則表示預測到之編碼區中屬於編碼核苷酸的比率。據此可整理成 2 × 2 兩向表(Fig. 1b)，並進行靈敏性(Sn)與專一性(Sp)的估算，公式如下：

$$Sn = \frac{TP}{TP + FN} \quad (1)$$

$$Sp = \frac{TP}{TP + FP} \quad (2)$$

2. 外顯子層次

針對核苷酸序列外顯子結構的預測，考慮三種可能預測結果(Fig. 1c)：正確外顯子(correct exon)、錯誤外顯子(wrong exon, WE)與遺漏外顯子(missing exon, ME)。針對外顯子為單位進行計算，亦以靈敏性表示實際的外顯子能被正確預測到的比率，而專一性表示預測到的外顯子中屬於實際外顯子的比率，公式如下：

$$Sn = \text{no. of correct exons / no. of actual exons} \quad (3)$$

$$Sp = \text{no. of correct exons / no. of predicted exons} \quad (4)$$

根據上述 3 與 4 式的評估方法，軟體必

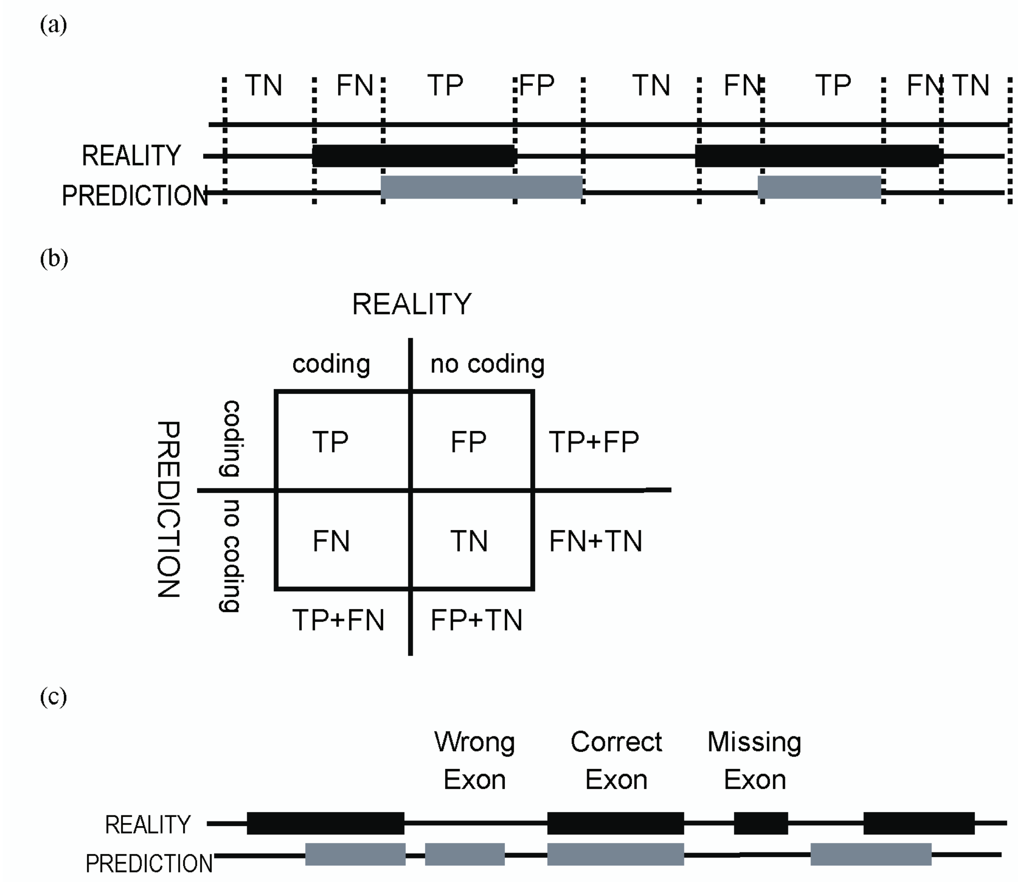

Fig. 1. Measures of prediction accuracy at the nucleotide and exon levels (modified from Burset and Guigó 1996).

須能正確預測到外顯子的位置與數目，才能獲得較高的預測結果，此種評估方式相較於核苷酸序列層次則更為嚴苛。因此，須再同時考量以預測遺漏外顯子的靈敏性與預測錯誤外顯子專一性的比率，始能對預測軟體的效率進行整體評估。公式如下：

Sn-ME = no. of missing exons/no. of actual exons (5)

Sp-WE = no. of wrong exons/no. of predicted exons (6)

四、程式撰寫與加值資料庫分析平台建立

由於本研究的序列材料是從網路上公開資料庫搜尋而得，該等資料庫的序列數量和內容都可能隨時更新，因此執行期間不定期的重新上線比對、下載、分析與整理資料是絕對必要的。為節省大量人力與時間，故本研究利用 Linux 作業環境建立自有之生物資訊分析平台，並以 Perl 語言(Scbwartz *et al.* 2006)撰寫相關程式，對 5 種軟體預測所得之外顯子位置與實際外顯子位置比較，自動計算出各基因序列的 TP、FP、TN、FN 值及靈敏性和專一性各項指標值，使基因預測軟體的評估變成自動化作業。同時，為使各軟體之預測結果間的差異性更容易辨識比較，亦以 Perl 語言的模組程式 Bioperl (http://www.bioperl.org/wiki/Main_Page)，設計一個基因模式化視覺平台，讓各軟體所預測出的序列結構能以圖形瀏覽方式呈現。

結果

就 5 種基因預測軟體對各 Xa 基因在核苷酸序列層次之預測準確度的靈敏性與專一性分析結果，如 Table 3 所示。GeneMark.hmm 軟體在所有 Xa 基因的 Sn 值表現，有 14 個達到 90%以上的預測準確度，而 Sp 值的表現除 Xa21F 為 97%外，其餘皆達到 100%的預測準確度，平均所有 Xa 基因的 Sn 與 Sp 值分別為 89%及 100%，顯示其以每個核苷酸為單位的預測準確度相當高。其次為 GENSCAN 軟體，Sn 值有 11 個達到 90%以上的預測準確度，而 Sp 值有 12 個達到 90%以上的預測準確度，平均所有 Xa 基因的 Sn 與 Sp 值分別為 86%及 89%；而 GlimmerM 軟體因為沒有預測到 Xa26d 基因，造成其平均所有 Xa 基因的 Sn 與 Sp 值各降為 77%及 92%，預測準確度因而略低於 GENSCAN。

針對正確預測到外顯子層次之預測準確度的 Sn 與 Sp 分析，結果示如 Table 4。綜觀 5 種預測軟體的表現，亦以 GeneMark.hmm 軟體之表現最佳，平均所有 Xa 基因的 Sn 與 Sp 分別為 44%及 44%，其中有 7 個基因的外顯子位置全部預測正確，但也有 9 個基因的外顯子位置預測有偏差而不完全正確。GeneMark.hmm 在平均所有 Xa 基因的 Sn-ME 及 Sp-WE 分別為 4%及 3%，顯示根據前述分析結果其預測外顯子位置的準確度雖然不高，但發生遺漏與錯誤外顯子的情形亦極少，故整體而言，GeneMark.hmm 的預測結果還算不錯；至於 GENSCAN 軟體在預測外顯子層次上，平均所有 Xa 基因的 Sn 及 Sp 各為 19%及 19%與 Sn-ME 及 Sp-WE 各為 7%與 20%，效果次於 GeneMark.hmm。綜合核苷酸與外顯子層次的預測結果(Table 3 及 Table 4)，皆以 GeneMark.hmm 的預測結果最佳，其次為 GENSCAN。

本研究為使基因預測結果除了常用的統計指標外，也能經由目視直接快速判讀，故另以 Bioperl 模組程式設計了一套適用於本研究的基因模式化視覺平台，將各軟體預測到之外顯子位置和實際外顯子位置，各用不同顏色的圖形方式標示以相互比較，其輸入介面與輸出結果示如 Fig. 2 和 Fig. 3。如此，當對某特定 DNA 序列進行預測而獲得結果後，即可透過 Fig. 2 之使用者介面輸入外顯子相關位置，而產生如 Fig. 3 之瀏覽畫面，讓預測結果之展現及比較上更一目了然。以 Xa26b 基因序列為例(Fig. 3)，其兩段外顯子的實際位置各在 1542~4476 bp 和 4582~4958 bp，倘使用 GeneMark.hmm 和 GENSCAN 則完全預測正確；GlimmerR 和 RiceHMM 能預測到正確的編碼區，但在第一個外顯子上仍有少數核苷酸沒有預測到；GlimmerM 則僅能完全預測到第一個外顯子，但卻無法預測到第二個外顯子。利用此種圖形瀏覽方式來呈現基因預測結果，可以馬上看出哪個預測軟體之效率最好，確實對基因預測軟體效益之評估指標(靈敏度及專一性)比較，可收相輔相成之效。

討論

本研究的評估結果以 GeneMark.hmm 軟體對於 Xa 基因的預測準確度最高，顯示 GeneMark.hmm 軟體的基因預測模式相較於其他預測軟體而言，已充分考量到此類型結構之基因序列。此預測軟體是 5 種軟體中最早開發的，而對於預測基因所採用之機率模式，亦根據編碼區與非編碼區的不同結構特性，分別採用 5 階 MM 與同質 MM 來進行基因預測(Borodovsky and McIninch 1993, Lukashin and Borodovsky 1998, Lomsadze *et al.* 2005)，此種預測策略有別於其他預測軟體所使用的單一機率模式，且更符合 DNA 序列中基因結構區域的分布型態。GENSCAN 是最常被用來做為基因預測軟體比較評估的對照軟體(Burest and Guigó 1996, Burge and Karlin 1997, Pedersen and Hein 2003)，雖從未經過水稻序列資料之訓練，但對水稻 Xa 基因的預測表現僅次於

Table 3. Prediction accuracy (%) at nucleotide level of five gene prediction programs on different Xa gene sequences.

	Xa1		Xa21									Xa26					Xa27		Mean
	1999	2004	A1	A2	B	B*	C	C*	D	E	F	a	b	b*	c	d	IR24	IRBB27	
GlimmerM																			
Sn	66	51	100	56	71	87	90	90	89	100	85	68	81	98	75	0	90	90	77
Sp	100	100	100	100	100	100	100	100	92	92	94	97	100	100	100	0	93	93	92
GlimmerR																			
Sn	98	75	96	31	71	100	65	65	92	79	85	11	83	89	0	22	90	100	70
Sp	100	100	100	94	100	100	98	98	100	97	94	100	100	99	0	79	80	100	91
RiceHMM																			
Sn	89	81	73	38	76	76	72	72	45	56	68	82	82	86	73	37	100	100	73
Sp	98	100	100	100	99	99	99	99	16	79	97	100	100	86	100	89	86	86	91
GeneMark.hmm																			
Sn	100	98	100	86	92	99	57	57	100	100	98	94	95	100	90	34	100	100	89
Sp	100	100	100	100	100	100	100	100	100	100	97	100	100	100	100	100	100	100	100
GENSCAN																			
Sn	61	98	87	89	99	100	52	74	84	91	97	100	95	100	97	37	90	90	86
Sp	100	100	98	96	100	100	87	91	23	84	97	100	100	100	100	61	80	80	89

Table 4. Prediction accuracy (%) at exon level of five gene prediction programs on different Xa genes sequences.

	Xa1		Xa21									Xa26					Xa27		Mean
	1999	2004	A1	A2	B	B*	C	C*	D	E	F	a	b	b*	c	d	IR24	IRBB27	
GlimmerM																			
Sn	0	0	100	0	0	50	0	0	0	0	0	0	0	50	0	0	0	0	11
Sp	0	0	100	0	0	50	0	0	0	0	0	0	0	33	0	0	0	0	10
Sn-ME	0	0	0	50	50	50	0	0	0	0	0	50	0	0	0	100	0	0	17
Sp-WE	0	0	0	0	50	50	0	0	0	75	67	0	0	0	0	100	0	0	19
GlimmerR																			
Sn	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	8
Sp	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	7
Sn-ME	0	0	0	50	0	0	50	50	0	0	0	50	50	50	100	67	0	0	26
Sp-WE	0	0	0	50	0	0	50	50	0	33	50	0	0	50	100	50	0	0	24
RiceHMM																			
Sn	0	0	50	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	6
Sp	0	20	20	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	4
Sn-ME	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	67	0	0	7
Sp-WE	0	0	0	0	0	0	20	20	80	57	20	0	0	25	0	33	50	50	20
GeneMark.hmm																			
Sn	100	0	100	0	0	50	0	0	100	100	0	0	50	100	0	0	100	100	44
Sp	100	0	100	0	0	33	0	0	100	100	0	0	50	100	0	0	100	100	44
Sn-ME	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	67	0	0	4
Sp-WE	0	0	0	0	0	33	0	0	0	0	25	0	0	0	0	0	0	0	3
GENSCAN																			
Sn	0	0	0	0	0	100	0	0	0	0	0	100	50	100	0	0	0	0	19
Sp	0	0	0	0	0	100	0	0	0	0	0	100	50	100	0	0	0	0	19
Sn-ME	0	0	0	0	0	0	0	0	0	0	0	0	0	0	67	67	0	0	7
Sp-WE	0	0	20	0	0	0	17	17	86	60	20	0	0	0	67	67	0	0	20

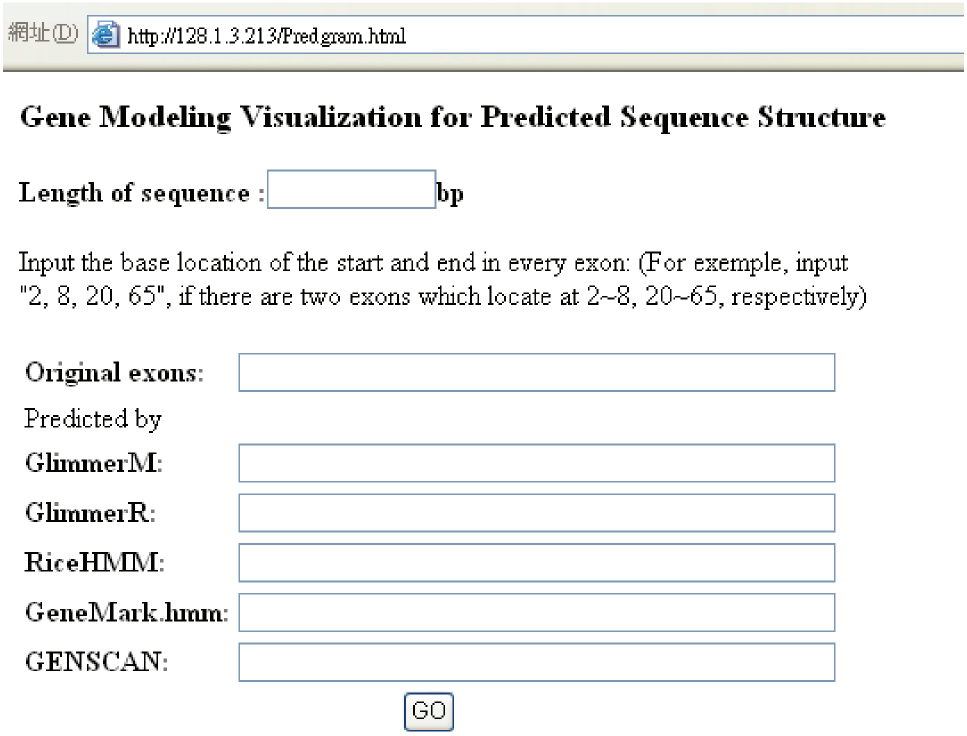

Fig. 2. Input interface of the gene modeling visualization platform.

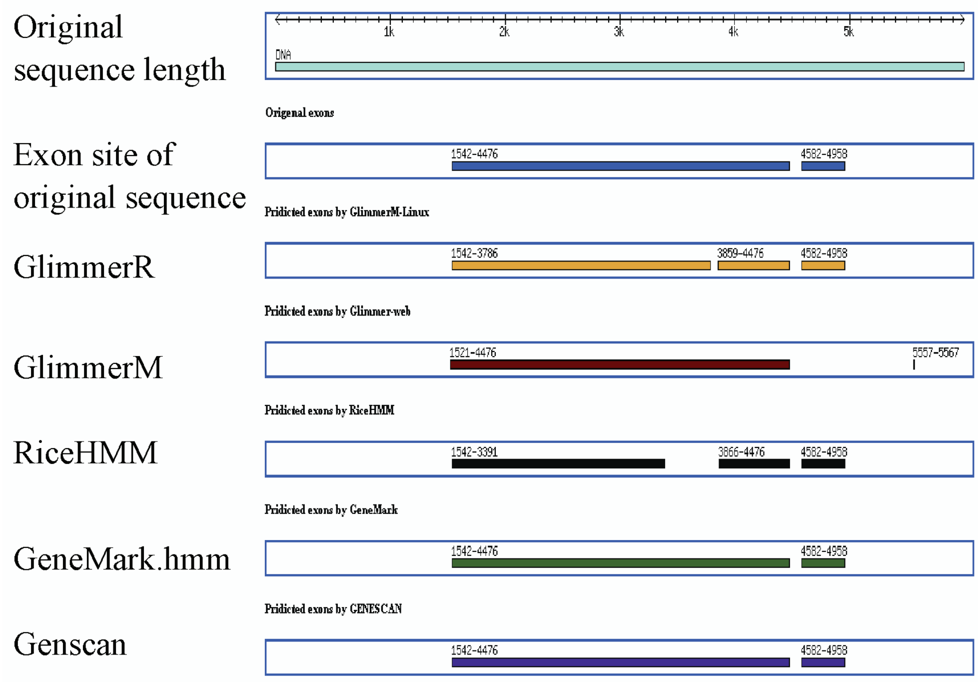

Fig. 3. Output display of the gene modeling visualization platform, taking Xa26b gene as an example.

GeneMark.hmm 軟體。另外，在 Linux PC 平台上所建立的 GlimmerM 軟體，對 Xa26d 序列並沒有預測到任何核苷酸，造成其平均所有 Xa 基因之 Sn 及 Sp 的值略低於 GENSCAN。由於 GlimmerM 軟體具有可建置於自有 Linux PC 平台的優勢，將來若能蒐集到適當的水稻基因序列資料，以該軟體提供的訓練模式加以訓練後，當更適合用來建立自有研究領域的基因預測分析平台。本研究亦發現其他 4 種預測軟體對 Xa26d 在核苷酸序列層次的預測結果也略低於其他基因 (Table 3)，同時 Table 4 中的 5 種預測軟體對 Xa26d 序列預測所獲得的 Sn-ME 及 Sp-WE 數值也都高於其他基因，顯示這幾種預測軟體對該 Xa26d 基因序列之預測準確度皆欠佳。檢視 Table 1 中各基因之 CDS 範圍(外顯子結構)，發現 Xa26d 序列的 3 個外顯子之間含有 2 個頗長的內含子(長度各為 6881 bp 和 2108 bp)，而其他基因序列之內含子皆在 1000 bp 以下，是否因為 Xa26d 存在此種特殊結構，以致於 GlimmerM 軟體無法預測到該基因，同時也造成其他預測軟體之預測準確度的降低，此值得未來再進一步探討。

本研究用來執行基因預測軟體評估的 18 個測試材料，經比對序列發表年份(Table 1)與預測軟體網站及文獻內容(Table 2)，推測應只有 Xa1-1999 與 Xa21 基因家族等共 10 個 DNA 序列，可能被納入 GlimmerR、RiceHMM 及 GeneMark.hmm 等 3 種預測軟體的訓練材料中，其餘 8 個 DNA 序列則為新的受測材料。雖然本研究限於材料來源之取得不易，致目前受測樣本較少，但所建立的基因預測軟體評估之相關分析流程及分析平台，仍有助於確認預測軟體在現有 Xa 基因之適用性。未來倘有較多新的 Xa 基因序列加入，更可評估出各預測軟體在此特殊結構之基因序列的優劣。

本研究以水稻白葉枯病抗性基因已解序的 DNA 序列為材料，找到適合目前資料型態的基因預測軟體，亦藉此建立了水稻白葉枯病抗性基因預測的分析流程及分析平台，有利於本研究團隊在後續的研究工作。未來將利用生物資訊與統計的方法，針對 Xa 基因序列結構進行深入剖析，找出 Xa 基因之 DNA 序列有別於其他水稻非 Xa 基因之 DNA 序列的特徵，配合水稻基因體序列資料庫做相似性的搜尋比對，希望能快速獲得大量符合 Xa 基因序列之結構特徵的序列片段；再以此既有之基因預測分析平台為基礎，找出與 Xa 基因序列之結構特徵有關的候選基因，以提供分子遺傳研究者進行基因定位實驗，節省傳統試驗所需的大量人力、物力與時間，加速新基因的發現。

誌謝

本研究承蒙行政院農業委員會農業生物技術國家型科技計畫支持(計畫編號：94 農科-5.2.1-農-C1(8))，特致謝忱。

引用文獻

Adhikari TB, A Shrestha, RC Basnyat, TW Mew (1999) Use of partial host resistance in the management of bacterial blight of rice. **Plant Disease** 83:896-901.

Birney E, R Durbin (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. **ISMB** 5:56-64.

Borodovsky M, J McIninch (1993) GeneMark: parallel gene recognition for both DNA strands. **Computers Chem**. 17:123-133.

Brejová B, DG Brown, M Li, T Vinař (2005) ExonHunter: a comprehensive approach to gene finding. **Bioinformatics** 21:i57-i65.

Burest M, R Guigó (1996) Evaluation of gene structure prediction programs. **Genomics** 34:353-367.

Burge C, S Karlin (1997) Prediction of complete gene structures in human genomic DNA. **J. Mol. Biol**. 268:78-94.

Burge CB, S Karlin (1998) Finding the genes in genomic DNA. **Curr. Opin. Stru. Biol**. 8: 346-354.

Delcher AL, D Harmon, S Kasif, O White, SL Salzberg (1999) Improved microbial gene identification with GLIMMER. **Nucleic Acids Res**. 27:4636-4641.

Goff SA, D Ricke, TH Lan, G Presting, R Wang, M Dunn, J Glazebrook, A Sessions, P Oeller, H Varma, et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). **Science** 296:92-100.

Gu K, B Yang, D tian, L Wu, D Wang, C Sreekala, F Yang, Z Chu, GL Wang, FF White, Z Yin (2005) R gene expression induced by a type-III effector triggers disease resistance in rice. **Nature** 435:1122-1125.

Henderson J, S Salzberg, K Fasman (1997) Finding genes in DNA with a hidden Markov model. **J. Comp. Biol**. 4:127-141.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. **Nature** 409:860-921.

Krogh A, M Brown, S Mian, K Sjölander, D Haussler (1994) Hidden Markov models in computational biology applications to protein modeling. **J. Mol. Biol**. 235:1501-1531.

Krogh A (1997) Two methods for improving performance of an HMM and their application for gene finding. **Proc. Intl. Conf. Intell. Syst. Mol. Biol**. 5:179-186.

Lomsadze A, V Ter-Hovhannisyan, Y Chernoff, M Borodovsky (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. **Nucleic Acids Res**. 33:6494-6506.

Lukashin AV, M Borodovsky (1998) GeneMark.hmm: new solutions for gene finding. **Nucleic Acids Res**. 26:1107-1115.

Mathé C, MF Sagot, T Schiex, P Rouzé (2002) Current methods of gene prediction, their strengths and weaknesses. **Nucleic Acids Res**. 30:4103-4117.

Pavy N, S Rombauts, P Déhais, C Mathé, DVV Ramana, P Leroy, P Rouzé (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. **Bioinformatics** 15:887-899.

Pedersen JS, J Hein (2003) Gene finding with a hidden Markov model of genome structure and evolution. **Bioinformatics** 19:219-227.

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. **Proc. IEEE** 77:257-285.

Rogic S, AK Mackworth, FBF Ouellette (2001) Evaluation of gene-finding programs on mammalian sequences. **Genome Res**. 11: 817-832.

Rogic S, FBF Ouellette, AK Mackworth (2002) Improving gene recognition accuracy by combining predictions from two gene-finding programs. **Bioinformatics** 18:1034-1045.

Sakata K, H Nagasaki, A Idonuma, K Waki, M Kise, T Sasaki (1999) A computer program for prediction of gene domain on rice genome sequence p.78. *In:* 2nd Georgia Tech. Intl. Conf. Bioinfor.:Sequence, Structure and Function. Nov. 11-14, 1999. Atlanta, Georgia, USA. (http://rgp.dna.affrc.go.jp/RiceHMM/)

Salzberg S, AL Delcher, S Kasif, O White (1998) Microbial gene identification using interpolated Markov models. **Nucleic Acids Res**. 26:544-548.

Scbwartz RL, T Phoenix, BD Foy (2006) Learning Perl. 4th ed. O'Reilly, USA. 316pp.

Song YL, GL Wang, LL Chen, HS Kim, LY Pi, T Holsten, J Gardner, B Wang, WX Zhai, LH Zhu, Z Fauquest, P Ronald (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. **Science** 270:1804-1806.

Staden R, AD McLachlan (1982) Codon preference and its use in identifying protein coding regions in long DNA sequences. **Nucleic Acids Res**. 10:141-156.

Stanke M, S Waack (2003) Gene prediction with a hidden Markov model and a new intron submodel. **Bioinformatics** 19:ii215-ii225.

Sun X, Y Cao, Z Yang, C Xu, X Li, S Wang, Q Zhang (2004) Xa26, a gene conferring resistance to *Xanthomonas oryzae* pv. *oryzae* in rice, encodes an LRR receptor kinase-like protein. **Plant J**. 37:517-527.

Sun X, Y Cao, S Wang (2006) Point mutations with positive selection were a major force during the

evolution of a receptor-kinase resistance gene family of rice. **Plant Physiol**. 140:998-1008.

Wang Z, Y Chen, Y Li (2004) A brief review of computational gene prediction methods. **Geno. Prot. Bioinfo**. 2:216-221.

Yeh RF, LP Lim, CB Burge (2001) Computational inference of homologous gene structures in the human genome. **Genome Res**. 11:803-816.

Yoshimura S, U Yamanouchi, Y Katayose, S Toki, ZX Wan, I Kono, N Kurata, M Yano, N Iwata, T Sasaki (1998) Expression of Xa1. A bacterial blight-resistance gene in rice, is induced by bacterial inoculation. **Genetics** 95:1663-1668.

Yu Y, T Rambo, J Currie, et al. (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. **Science** 300:1566-1569.

－編輯：楊純明